Iran. J. Insur. Res., 12(1): 15-26, Winter 2023, Serial #43

Iranian Journal of Insurance Research
(IJIR)

Homepage: https://ijir.irc.ac.ir/?lang=en

Insurance Research Center (IRC)

ORIGINAL RESEARCH PAPER

Classification of life insurance customer point of views based on text mining algorithms

A. Aminpour[1,*], M. Rabiei[2]

[1] *Department of Industrial Engineering, Faculty of Engineering, Eyvanekey University, Eyvanekey, Iran*

[2] *Department of Electrical and Computer Engineering, Faculty of Engineering, Eyvanekey University, Eyvanekey, Iran*

ARTICLE INFO

Article History:
Received 31 May 2022
Revised 05 September 2022
Accepted 25 October 2022

Keywords:
Knowledge Discovery
Life Insurance
Machine Learning
Textual Data
Text Mining

ABSTRACT

BACKGROUND AND OBJECTIVES: In recent years, the insurance industry has grown significantly and different companies started working with various services in this field. Since successful marketing is one of the main goals of insurance companies, it is very important to find people who are likely to want to use life insurance services. This achievement can lead to better management of capital and costs. The main objective of this research is to classify the views of life insurance customers of an insurance company based on text mining algorithms, so that this classification can be used as a basis for predicting potential customers in the future. Anticipating this category of customers. In that case, we will be able to adopt a suitable marketing strategy to sell our services.

METHODS: In this research, we have analyzed a textual dataset, including life insurance customer's opinions. Despite the growing volume of this type of data, there are applicable tools for organizing, retrieving and discovering useful knowledge from them. In this regard, this research has been carried out on text processing techniques. These techniques seek useful information from unstructured textual data using pattern recognition and discovery. In this article, the views of customers related to life insurance have been examined as an independent issue. The main goal is to categorize these comments into positive and negative categories based on text mining algorithms. To achieve this objective, for the first time in the insurance industry, four different machine algorithms are used in line with text mining of policyholders' points of view.

FINDINGS: According to the techniques used in this research and the obtained results, it can be said that the support vector machine algorithm has the highest prediction accuracy criterion with 73% compared to other algorithms used in this research. At the same time, most of the insurance policyholders have also expressed a positive opinion about the services received, and this means that most of the customers using the mentioned services were satisfied with the company.

CONCLUSION: The majority of insured would like to keep this insurance service in their shopping basket in the future. Therefore, company managers can find their potential customers from among these people and plan to sell their services to them. By adopting this type of marketing strategy, managers can reduce the costs of their company and reduce the price of their services by saving marketing costs. It is natural that one of the important goals of any company is to earn more profit, and this will not be possible unless it maintains its customers by offering optimal prices and increases them day by day. Achieving this depends on our cost understanding, price acceptance, consumer satisfaction and strategic marketing actions. By exploiting the results of this research, it is possible to achieve a suitable marketing strategy for determining the price of insurance services. Determining the optimal price of insurance premium is considered a competitive advantage for companies. The price in all industries is subject to the law of supply and demand. Since getting the best price is one of the top priorities of insurance customers, even a small percentage change in premium prices will cause many customers to switch insurers. Therefore, optimal pricing can be very effective in increasing insurance profits.

*Corresponding Author:
Email: *Aminpour_A@eyc.ac.ir*
Phone: +9821 44337368
ORCID: 0000-0003-0243-4067

DOI: 10.22056/ijir.2023.01.02

نشریه علمی پژوهشنامه بیمه دوره ۱۲، شماره ۱، زمستان ۱۴۰۱، ص ۱۵-۲۶

نشریه علمی پژوهشنامه بیمه

سایت نشریه: https://ijir.irc.ac.ir/?lang=fa

مقاله علمی

دسته‌بندی نظرات خریداران بیمه زندگی بر اساس الگوریتم‌های متن‌کاوی

علیرضا امین پور[1*]، محمد ربیعی[2]

[1] گروه مهندسی صنایع، دانشکده فنی و مهندسی، دانشگاه ایوان کی، ایوان کی، ایران

[2] گروه مهندسی برق و کامپیوتر، دانشکده فنی و مهندسی، دانشگاه ایوان کی، ایوان کی، ایران

اطلاعات مقاله

تاریخ های مقاله:

تاریخ دریافت: ۱۰ خرداد ۱۴۰۱

تاریخ داوری: ۱۴ شهریور ۱۴۰۱

تاریخ پذیرش: ۰۳ آبان ۱۴۰۱

کلمات کلیدی:

بیمه زندگی

پردازش متن

داده متنی

کشف دانش

یادگیری ماشین

*نویسنده مسئول:

ایمیل: Aminpour_A@eyc.ac.ir

تلفن: ۴۴۳۳۷۳۶۸ ۹۸۲۱+

ORCID: 0000-0003-0243-4067

چکیده:

پیشینه و اهداف: در سال‌های اخیر صنعت بیمه رشدی چشمگیر داشته است و شرکت‌های مختلف با خدمات گوناگون پا به عرصه گذاشته‌اند. بازاریابی موفق یکی از اهداف اصلی شرکت‌های بیمه است؛ پیداکردن افرادی که احتمال می‌رود از خدمات بیمه استفاده کنند، بسیار مهم است و منجر به مدیریت هرچه بهتر سرمایه و هزینه‌ها می‌شود. هدف اصلی این پژوهش، دسته‌بندی نظرات خریداران بیمه زندگی یک شرکت بیمه ای بر اساس الگوریتم‌های متن‌کاوی است تا بتوان از این دسته‌بندی به عنوان مبنایی برای پیش‌بینی مشتریان احتمالی آتی استفاده کنیم. با پیش‌بینی این دسته از مشتریان می‌توانیم استراتژی بازاریابی مناسبی برای فروش خدمات خود اتخاذ کنیم.

روش‌شناسی: در این پژوهش به بررسی یک مجموعه داده متنی شامل نظرات خریداران بیمه زندگی پرداخته‌ایم، چرا که با وجود رشد روز افزون حجم این دسته از داده‌ها، وجود ابزارهایی جهت سازماندهی، بازیابی و کشف دانش مفید از آنها ضروری است. در همین راستا، تاکنون تحقیقات گسترده‌ای روی تکنیک‌های پردازش متن صورت گرفته است. این تکنیک‌ها، با استفاده از شناسایی و کشف الگوها، به دنبال استخراج اطلاعات مفید از داده‌های متنی بدون ساختار هستند. در این مقاله نظرات خریداران بیمه زندگی، به صورت یک مساله مستقل مورد بررسی قرار گرفته است. هدف اصلی، دسته‌بندی این نظرات بر اساس الگوریتم‌های متن‌کاوی به دو دسته مثبت و منفی است. برای رسیدن به این هدف، برای اولین بار در صنعت بیمه از چهار الگوریتم مختلف یادگیری ماشین برای متن‌کاوی نظرات بیمه‌گذاران استفاده شده است.

یافته‌ها: با توجه به نتایج حاصله از تکنیک‌های به کار رفته در این پژوهش می‌توان گفت که الگوریتم ماشین‌بردار پشتیبان با میزان ۷۳ درصد، بیشترین میزان معیار دقت پیش‌بینی را در بین سایر الگوریتم‌های مورد استفاده در این پژوهش داشته است. در ضمن اکثریت بیمه‌گذاران نیز نظر مثبتی در ارتباط با خدمات دریافتی داشته‌اند و این بدان معناست که اکثر مشتریان استفاده کننده از خدمات، از شرکت راضی هستند.

نتیجه‌گیری: اکثریت بیمه‌گذاران مایلند در آینده نیز این خدمت بیمه‌ای را در سبد خرید خود داشته باشند. لذا مسئولین شرکت می‌توانند، مشتریان احتمالی خود را از میان این افراد پیدا و برای فروش خدمات خود بر روی آنها سرمایه‌گذاری کنند. با این استراتژی بازاریابی، مدیران می‌توانند هزینه‌های شرکت را کاهش داده و با صرفه‌جویی در هزینه از این راه، قیمت خدمات خود را کاهش دهند. همه ما می‌دانیم که هدف هر شرکتی تعیین قیمت برای به حداکثررساندن سود است که به آن قیمت بهینه نیز گفته می‌شود. تعیین قیمت بهینه به درک هزینه‌ها، کشش قیمت، ترجیحات مصرف‌کننده و اقدامات استراتژیک بازاریابی ما بستگی دارد. با این نتایج می‌توانیم استراتژی بازاریابی مناسب خود را انتخاب کنیم. زیرا تعیین یک قیمت حق بیمه بهینه یک مزیت رقابتی برای شرکت‌ها ایجاد می‌کند. مانند هر صنعت دیگری، قیمت تابع قانون عرضه و تقاضا است. از آنجایی که دریافت بهترین قیمت جزو اولویت‌های اصلی مشتریان بیمه است، حتی درصد کمی تغییر در قیمت حق بیمه باعث می‌شود بسیاری از مشتریان بیمه‌گران خود را تغییر دهند. بنابراین، قیمت‌گذاری بهینه در بخش بیمه، حداکثر سود را ممکن می‌سازد.

DOI: 10.22056/ijir.2023.01.02

مقدمه

در جهان کنونی که امکان تولید انبوه کالا و خدمات، زمینهٔ لازم افزایش عرضه نسبت به تقاضا را فراهم آورده است، برای کسب و کارها راهی جز تعیین مشتریان بالقوهٔ احتمالی باقی نمانده و دیگر نمی‌توان حیطهٔ بازار و عرضه را با ابزارهای محدود گذشته تعریف کرد. (Berson *et al.*, 1999) با ظهور اقتصاد رقابتی، مفاهیمی چون مشتری مداری و مشتریان راضی پایه و اساس کسب و کار تلقی شده و سازمانی که بدان بی‌توجه باشد از صحنهٔ بازار حذف خواهد گردید.

شرکت‌های بیمه‌ای یکی از بزرگترین کسب و کارها و جزو نهادهای مؤثر در وضعیت اقتصادی هر کشوری محسوب می‌گردند و رشد صنعت بیمه بیانگر توسعه‌یافتگی و افزایش پس‌اندازهای مالی در هر کشور است. شرکت‌های بیمه با کنترل کامل چرخهٔ بازاریابی، فروش و خدمات در تمام رشته‌های بیمه‌ای و رسیدگی دقیق به تمام درخواست‌های مشتریان، می‌توانند گام مؤثری در جهت حفظ مشتریان بردارند. از این‌رو، شناسایی مؤلفه‌های اصیل مؤثر بر رضایت مندی مشتریان باید در اولویت برنامه‌های شرکت‌های بیمه‌ای قرار گیرد(Motarjem and Niakan, 2020) . «با پیشرفت سریع فناوری اطلاعات، میزان اطلاعات ذخیره شده در پایگاه‌های دادهٔ بیمه نیز به سرعت در حال افزایش است. این پایگاه‌های دادهٔ بزرگ، منبع با ارزش و بالقوهٔ اطلاعات این کسب و کار را تشکیل می‌دهند» (Hsia *et al.*, 2000).

در گذشته عموماً استخراج اطلاعات مفید از داده های ثبت شده، به صورت دستی و بر عهدهٔ تحلیل گران بوده است. با توجه به این که تجزیه و تحلیل دستی داده ها بسیار کند و گران بود و هر روز بر پیچیدگی و حجم داده ها افزوده می شد، تحلیل به سمت تحلیل های غیر مستقیم خودکار و استفاده از روش های رایانه ای حرکت کرده است (Vali Mohammadi and Shekarchi, 2010). نیاز مبرم مبنی بر استفاده از فناوری های جدید و ابزارهای خودکار سبب شد تا امروزه به صورت هوشمند حجم زیاد داده را به اطلاعات و دانش تبدیل کنند (Hajiheydari *et al.*, 2011).

در حقیقت، بخش اعظمی از داده‌های در دسترس امروز تنها طی چند سال اخیر تولید شده اند. لذا، دنیای امروز دنیای شکوفایی داده‌هایی با حجم، سرعت و تنوع بالا است که تحت عنوان کلان داده‌ها شناخته می‌شوند .(Sagiroglu and Sinanc, 2013) در صورتی که این داده‌ها به خوبی تحلیل شوند و ارزیابی‌های صحیحی نسبت به آن‌ها صورت گیرد، می‌تواند ابزار مهمی جهت پیش‌بینی مشتریان احتمالی آتی بیمه قلمداد شود. استفادهٔ بهینه از پایگاه‌های داده در صورتی امکان پذیر خواهد بود که از ابزارهای کارا و استاندارد در کنار یک برنامهٔ اصولی و آینده نگر برای تبدیل داده به اطلاعات کمک بجوییم. بنابراین، امروزه با فقدان یا کمبود اطلاعات مواجه نیستیم، بلکه آنچه از اهمیت بسزایی برخوردار است استفاده از روش‌های مناسب و استاندارد جهت نگهداری، به روز کردن، در دسترس قرار دادن و نهایتاً کشف دانش‌های جدید از انبوه اطلاعات موجود است (Zaresaadabadi, 2014). شرکت‌های بیمه با کنترل کامل چرخهٔ بازاریابی، فروش و خدمات در تمام رشته‌های بیمه‌ای و رسیدگی دقیق به تمام درخواست‌های مشتریان می‌توانند گام مؤثری در جهت حفظ مشتریان بردارند. از این‌رو، شناسایی مؤلفه‌های اصیل مؤثر بر رضایت مندی مشتریان باید در اولویت برنامه‌های شرکت‌های بیمه‌ای قرار گیرد (Motarjem and Niakan, 2020). از سوی دیگر، در بازارهای رقابتی به علت وجود بازیگران متعدد، فشار رقابتی برای کاهش قیمت‌ها وجود دارد که صنعت بیمهٔ ایران نیز به علت افزایش قابل توجه شرکت‌ها ناشی از خصوصی‌سازی‌های آغاز شده از سال ۱۳۸۲ از این قاعده مستثنی نیست(Manteghipour and Alaei, 2022). لذا، استخراج دانش از کلان‌داده‌های متنی، صوتی و تصویری می‌تواند علاوه بر شناسایی نقاط ضعف شرکت و عوامل ایجاد نارضایتی در مشتریان، به شناسایی مشتریان راضی و تعیین جامعهٔ هدف برای تعیین استراتژی‌های بازاریابی کمک شایانی نماید. در چنین شرایطی است که باید از رشد فناوری برای استفادهٔ مؤثر از این دانش بالقوه سود جست و داده کاوی یک جواب مناسب برای استخراج این ثروت است (Oliaei *et al.*, 2018).

خاطر نشان می‌شود با توسعهٔ پردازش زبان طبیعی و یادگیری ماشین، راه برای تحلیل سریع داده‌های انبوه هموارتر شده است. پیشرفت های رایانه‌ای اجازهٔ تقسیم بندی و تمرکز بهتر بر روی رشته‌های مختلف بیمه ای را فراهم آورده و واحدهای تجاری را بیش از پیش قادر به توسعهٔ مدل‌های تجاری، طراحی تولیدات جدید و تقسیم بندی بیشتر مشتریان گوناگون، ساخته است. از این رو، شرکت های بیمه با به کارگیری پیشرفت‌های به دست آمده، با تأکیدی روز افزون از حالت معمولی و سنتی ارائهٔ خدمات بیمه ای، به خدمات جدید با تقسیم بندی هدفمند مشتریان روی آورده اند. اطلاعات مربوط به بیمه گذاران و داده‌های تجاری در بانک های اطلاعاتی ذخیره و بایگانی شده و در سیستم داده کاوی مورد بررسی قرار می‌گیرند(Gharakhani and Abolghasemi, 2011).

به طور کلی، پژوهش‌ها نشان می دهند تجربهٔ مشتریان از مصرف محصولات یا دریافت خدمات، عامل اصلی موفقیت در بازاریابی جدید است (Yoon and Lee, 2017). این موضوع به ویژه در حوزه خدمات و از جمله بیمه از اهمیت بسیار زیادی برخوردار است و شرکت‌ها باید در جهت ایجاد روابط بلندمدت و رضایت مشتریان تلاش کنند (Homburg *et al.*, 2017). در برخی کارهای پیشین که در این زمینه انجام شده است، علی رغم هزینهٔ زیاد برای طراحی سیستم، به دلیل وجود برخی از اشکالات در طراحی مدل، نتیجهٔ چشم‌گیری حاصل نشده است. یکی دیگر از علل عدم حصول موفقیت چشم‌گیر می‌تواند فقر منبع اطلاعاتی باشد. در کار پیش رو، سعی بر آن شده است با هدف دسته‌بندی خریداران بیمهٔ زندگی یک شرکت بیمه ای، به دو دستهٔ مثبت و منفی و به تبع آن شناسایی مشتریان با رویکرد مثبت نسبت به شرکت و تعیین آنها به عنوان مشتریان بالقوهٔ آتی، به تحلیل پیام‌های متنی این دسته از مشتریان پرداخته شود. در این تحقیق، داده‌های جمع‌آوری شده ابتدا توسط روش‌های پیش پردازش مورد غربال قرار می‌گیرند تا ناخالصی‌ها و داده‌های ناقص آن حذف گردد و سپس سعی می‌کنیم با استفاده از تکنیک‌های متن کاوی، مشتریان را به دو دستهٔ مثبت و منفی (از نقطه نظر میزان رضایت) دسته‌بندی نماییم. بدین ترتیب، دستهٔ مثبت می‌تواند مشتریان بالقوهٔ

آتی ما را مشخص کرده و ما این امکان را داریم که استراتژی بازاریابی خود را بر اساس تعداد و ویژگی‌های این دسته طراحی نماییم. امروزه در صنعت‌بیمه عوامل مهمی‌همچون عدم ایجاد، حفظ و توسعهٔ روابط تجاری سودمند با مشتریان، عدم استفاده از روش های نوین مشتری مداری، عدم جایگاه یابی مناسب برای برخی از بیمه‌نامه‌ها و در نهایت عدم فروش این بیمه نامه‌ها مطرح می‌باشد که با استفاده از بازاریابی مناسب می توان این نقاط ضعف را به نقاط قوت تبدیل نمود (Amarasinghe *et al.*, 2021). نوآوری این پژوهش استفاده از چهار الگوریتم مختلف یادگیری ماشین برای متن کاوی نظرات بیمه‌گذاران است که برای اولین بار در صنعت بیمهٔ زندگی ایران صورت می‌پذیرد. در این مقاله، پس از مبانی نظری و مروری بر پیشینهٔ پژوهش به معرفی داده‌های مورد بررسی پرداخته و روش پیاده سازی تحقیق را توضیح می‌دهیم. پس از آن، به روش‌های ارزیابی مدل پرداخته و نتایج را تشریح می‌نماییم. درنهایت، پیشنهادهایی برای تحقیقات آتی در این زمینه ارائه شده است.

مبانی نظری و مروری بر پیشینه پژوهش

داده کاوی در صنعت بیمه می‌تواند به شرکت‌ها در جهت کسب مزیت رقابتی کمک شایانی نماید. به طور مثال، با به کارگیری تکنیک‌های داده کاوی، شرکت‌ها می‌توانند با استفاده از داده‌ها در مورد الگوی خرید مشتری و رفتار مشتری به کشف دانش بپردازند. همچنین داده کاوی در درک بیشتر از کسب و کار برای کمک به کاهش تقلب، ارتقای سطح بیمه‌گری و بالا بردن مدیریت ریسک، ابزار مناسب و مؤثری ارائه می‌کند (Cheo Yeo *et al.*, 2001). به طور کلی پژوهش‌ها در زمینهٔ داده کاوی در صنعت بیمه به سه دسته تقسیم می‌شوند که شامل تحلیل ریسک و برآورد آن، کشف تخلفات و مدیریت ارتباط با مشتری می‌باشد.

در حوزهٔ مدیریت ارتباط با مشتری که پژوهش حاضر نیز در این حوزه قرار می‌گیرد، از فنون داده کاوی برای گروه‌بندی مشتریان و تحلیل الگوهای رفتاری مشتریان استفاده می‌شود. در این مقالات که در آنها بیشتر از روش‌های خوشه‌بندی و دسته‌بندی استفاده می‌گردد، سعی شده است تا با استفاده از روش‌های خوشه‌بندی ابتدا گروه‌های لازم ایجاد گردد، سپس با استفاده از روش‌های دسته‌بندی، هر کدام از مشتریان بر اساس ویژگی‌هایشان در هر یک از این گروه‌ها قرار گیرند.

Momeni and Ghodousi (2016) در مقاله‌ای تحت عنوان «سیستم های پشتیبانی از تصمیم و هوش کسب و کار و طراحی و پیاده سازی یک سیستم تصمیم یار (مطالعهٔ موردی در صنعت بیمه)» به بررسی نقش فناوری‌های کامپیوتری برای پشتیبانی از فرایندهای مدیریتی، به ویژه تصمیم‌گیری پرداخته اند. این مقاله شامل کاربرد روش‌های داده کاوی برای پیاده سازی هوش کسب و کار با بررسی روش‌ها، ابزارها و فرایندها و کاربرد شبکه‌های عصبی مصنوعی در تصمیم سازی و فرایندهایETL می‌شود. در این مقاله به عنوان نمونه یک سیستم تصمیم یار برای صنعت بیمه با استفاده از روش استنتاج مبتنی بر مورد پیاده سازی شده است.

Izadparast *et al.* (2022) در مقاله ای تحت عنوان «ارائهٔ چهارچوب برای پیش بینی سطح خسارت مشتریان بیمهٔ بدنهٔ اتومبیل با استفاده از راهکار داده کاوی» با استفاده از روش های داده کاوی، چهارچوبی را برای شناسایی مشتریان (با تمرکز بر مشتریان بیمهٔ بدنهٔ اتومبیل) ارائه کرده اند. در این پژوهش به منظور کاهش خطاها و محدودیت‌ها، از دو روش درخت تصمیم و خوشه بندی به صورت مکمل استفاده شده است. ابتدا با استفاده از تکنیک خوشه بندی، مشتریان بر اساس ویژگی هایشان خوشه بندی شده و سطح خسارت هرکدام از این خوشه‌ها محاسبه شده است. سپس مشتریان آتی بر اساس ویژگی هایشان و تکنیک درخت تصمیم در یکی از این خوشه‌ها دسته بندی شده اند. دسته ای که مشتری در آن قرار گرفته معرف سطح خطرپذیری اوست. با استفاده از این معیار و نوع بیمه نامهٔ مشتری، پیش بینی میزان خسارت وی میسر شده است. در نهایت، برای تعیین صحت نتایج، مدل ها ارزیابی و با یکدیگر مقایسه شده اند. در این پژوهش، تکنیک درخت تصمیم برای این منظور نتایج بهتری را حاصل نموده است.

Rezaei navaei and Koosha (2017) در مقاله‌ای تحت عنوان «به کارگیری و ارزیابی تکنیک‌های داده کاوی جهت پیش بینی روی گردانی مشتری در صنعت بیمه»، از تکنیک‌های شناخته شدهٔ دسته بندی داده کاوی برای پیش بینی روی گردانی مشتری در صنعت بیمه استفاده کرده اند. در این مقاله برای نخستین بار پیش بینی روی گردانی مشتری در یک سازمان بیمه ای با استفاده از یکی از رویکردهای مبتنی بر تحقیق در عملیات دسته‌بندی، یعنی تکنیک ماشین بردار پشتیبان انجام می‌شود. در این مقاله نخست از الگوریتم ژنتیک برای انتخاب مشخصه‌های تاثیرگذار استفاده شده است و سپس بعد از مدل سازی مسئله، پارامترهای مدل ماشین بردار پشتیبان با استفاده از دو روش جستجوی شبکه و اعتبار سنجی متقابل K لایه بهینه می‌شوند. در ادامه، عملکرد پیش بینی روش ماشین بردار پشتیبان با روش‌های درخت تصمیم شبکه‌های عصبی، رگرسیون لجستیک، جنگل تصادفی، دسته بندی کنندهٔ بیزی و K نزدیک ترین همسایگی مقایسه شده است. یافته‌های این تحقیق نشان می‌دهد که روش ماشین بردار پشتیبان از عملکرد بالاتری نسبت به سایر روش‌ها برخوردار است. در مدل پیشنهادی مبتنی بر این روش مشخصه‌های سابقهٔ خرید نحوهٔ آشنایی با سازمان و تمایل به خرید به عنوان مشخصه‌های اصلی پیش‌بینی‌کنندهٔ روی گردانی مشتری شناسایی شده اند.

Jun mei *et al.* (2015) برای حل مشکل عدم توازن توزیع داده‌ها و بهبود پیش‌بینی مشتریان با ارزش در صنعت بیمه، از الگوریتم جنگل تصادفی بهبود یافته استفاده کرده و پیش‌بینی می‌کنند که این عملکرد منجر به کاهش مشتریان از دست رفته خواهد شد. آنها با استفاده از نرم افزار CTREE و در نظر گرفتن سه دسته A و B وC که نشان دهندهٔ بیمهٔ زندگی، بیمهٔ حوادث خانواده و بیمهٔ حوادث فردی می‌باشد و با در دست داشتن داده‌ای مشتمل بر ۱۰۰۰ بیمه گذار در هر دسته، خطای دسته‌بندی اشتباه در داده‌های آموزشی را ۰،۰۷ درصد و در داده‌های آزمایشی ۰،۳۴ درصد به دست آوردند.

Senousy *et al.* (2018) در پژوهشی به عنوان «بررسی روند اخیر تجزیه و تحلیل کلان داده‌ها برای دستیابی به مدل‌های اصلاح شدهٔ صنعت بیمه» به نقش کلان داده‌ها و داده کاوی آنها در طراحی و تعیین مدل‌های اصلاح شده در صنعت بیمه می‌پردازند. در این پژوهش عنوان شده است که هدف از تجزیه و تحلیل حجم عظیمی از اطلاعات و داده‌های بزرگ در صنایع بیمه ای، بهبود دقت در پیش بینی، تعیین ریسک و شناخت تجربهٔ مشتری است. در این پژوهش از ابزارهای یادگیری ماشین و ابزارهای یادگیری عمیق و همچنین کتابخانهٔ یادگیری ماشین به عنوان یکی از برجسته ترین سیستم عامل‌ها برای تجزیه و تحلیل کلان داده‌ها برای مدل سازی استفاده شده است.

متن کاوی یکی از انواع داده کاوی است. متن کاوی از اطلاعات متنی غیر ساخت یافته استفاده کرده و آن را برای کشف ساختار و معناهای ضمنی پنهان در متن بررسی می کند (Boroun *et al.*, 2015). به عبارت دیگر، می‌توان گفت که متن کاوی نوعی داده کاوی بر روی داده‌های متنی است. ولی هدف، تکنیک‌ها و فرایند آن کمی متفاوت از داده کاوی است. از زمان ظهور این تکنیک، مدل‌های بسیاری ابداع، توسعه و مورد استفاده قرار گرفته‌اند. Turney (2002) در پژوهش خود، کلمات مورد استفاده در ۱۷۰۰۰ جملهٔ متنی را با استفاده از برچسب زن نقش کلمات مشخص کرد و سپس برای جهت‌یابی معنایی عبارات اسمی، از معیار PMI استفاده نمود که این معیار بیان‌کنندهٔ احتمال استفادهٔ همزمان دو کلمه در یک پیکرهٔ متنی است. اطلاعات متقابل بین عبارت داده شده و کلمهٔ«عالی» منهای اطلاعات متقابل بین عبارت داده شده و کلمهٔ «ضعیف» محاسبه گردیده، سپس توسط جهت معنایی عبارات شامل صفت، طبقه‌بندی انجام گرفته است. خروجی این سیستم به صورت مثبت یا منفی برای جملات در مقاله ارائه شده است.

Palakshappa and Patil (2022) در تحقیق خود با عنوان «طراحی مدل RFM برای پیش بینی رفتار خرید مشتری با استفاده از الگوریتم نزدیکترین همسایه در صنعت بیمه»، عنوان کرده است که هدف از این مطالعه استفاده از هوش تجاری در شناسایی مشتریان بالقوه با تهیهٔ اطلاعات مناسب و به موقع برای اشخاص تجاری در صنعت بیمه است. داده‌های ارائه شده مبتنی بر مطالعهٔ سیستماتیک و کاربردهای علمی در تجزیه و تحلیل تاریخ فروش و رفتار خرید مشتریان می‌باشد. در این پژوهش، به منظور اجرا و به کارگیری رویکرد علمی با استفاده از الگوریتم نزدیکترین همسایه، مجموعه داده‌های فروش خدمات بیمه ای مورد تجزیه و تحلیل قرار گرفته است. مقادیر مجموعه داده‌ها و پارامترهای مربوط به مدت زمان خاص فعالیت‌های تجاری، درک سازمان یافته ای از الگوهای خرید و رفتار مشتری در مناطق مختلف ارائه می‌دهد. این مطالعه با استفاده از مدل RFM انجام شده است و از اصول تقسیم بندی مجموعه داده‌ها با استفاده از الگوریتم نزدیکترین همسایه استفاده شده است.

Oshini and Caldera (2013) با تمرکز بر اجرای تکنیک‌های حفظ مشتری به موضوع داده کاوی در بیمه‌های زندگی پرداختند. آنها به این نتیجه رسیدند که اجرای تکنیک‌های داده کاوی در حوزهٔ بیمهٔ زندگی به راحتی می‌تواند از ریزش بیمه‌گذاران جلوگیری کند.

روش‌شناسی پژوهش

پژوهش حاضر از نوع داده محور است که با استفاده از فرایند استاندارد داده کاوی در صنعت انجام شده است. این فرایند شامل مراحل درک مسئله، درک داده‌ها، آماده‌سازی داده‌ها، مدل سازی، ارزیابی نتایج و به کارگیری مدل است (Hajiheydari *et al.*, 2011). روش پیشنهادی در این پژوهش، بر اساس مراحل نشان داده شده در شکل ۱، به شرح ذیل پیاده‌سازی شده است:

پیش پردازش

این مرحله شامل ریشه‌یابی و حذف کلمات ایستا است. روش‌های پیش پردازش متن از آن جهت حائز اهمیت هستند که ابزارهای لازم را برای تبدیل متن از زبان طبیعی به فرمت قابل خواندن توسط ماشین فراهم می‌کنند. در مرحلهٔ پیش پردازش، با استفاده از کتابخانه‌های موجود در زبان برنامه نویسی پایتون همانند Hazm به نرمال‌سازی داده‌ها پرداخته شده است.

روش FT-FDI

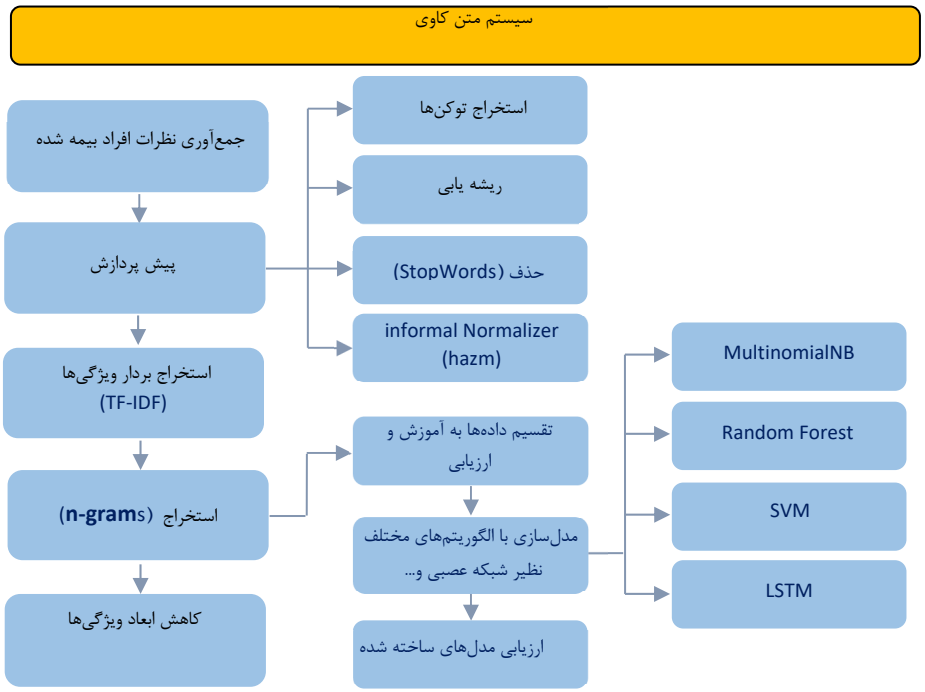
این روش یک روش آمار عددی است که میزان اهمیت یک کلمه نسبت به یک سند در مجموعه‌ای از اسناد را نشان می‌دهد و در واقع هدف آن، نشان دادن اهمیت کلمه در متن است (Farokhseir and Esmaeelpour, 2015). مقدار TF- IDF به تناسب تعداد تکرار کلمه در سند افزایش می‌یابند و توسط تعداد اسنادی که در مجموعه هستند و شامل کلمه نیز می‌باشند، متعادل می‌شود؛ به این معنی که اگر کلمه‌ای در بسیاری از متون ظاهر شود، احتمالاً کلمه‌ای متداول است و ارزش چندانی در ارزیابی متن ندارد.

استخراج n-smarg

منظور از n-grams در فرایندهای متن کاوی، دنباله‌ای به هم پیوسته از کلمات یا لغات می‌باشد(Asghari, 2012) . به عنوان مثال 3-grams به معنی دنباله‌های سه تایی از کلمات یا حروف می‌باشد. Bigrams به معنی دو کلمه است و کلمات موجود در یک جمله را به صورت دوبه دو در کنار یکدیگر قرار می‌دهد و مدل trigrams کلمات موجود در یک جمله را سه به سه در کنار یکدیگر قرار می‌دهد. n-grams به سادگی همهٔ ترکیب‌های مهم کلمات مجاور یا حروف با طول n را که در یک متن یافت می‌شود، نمایش می‌دهد. یکی از امکانات مهمی که n-grams ارائه می‌دهد این است که ساختار زبان را از نظر آماری ضبط می‌کند. طول مطلوب n به کاربرد و زمینهٔ استفادهٔ آن بستگی دارد. اگر n-grams بسیار کوتاه باشد ممکن است تفاوت مهمی را ثبت نکند و اگر خیلی طولانی باشد فقط اطلاعات کلی را به تصویر می‌کشد.

کاهش ابعاد با روش ACP

با اعمال این روش، متغیرهای ورودی اولیه به مولفه‌های جدید بدون همبستگی تبدیل می‌شوند، به طوری که مؤلفه‌های ایجاد شده،



شکل ۱: روش پیاده‌سازی

ترکیبی خطی از متغیرهای ورودی‌اند (Babania *et al.*, 2020). تحلیل مؤلفه‌های اصلی یکی از کاربردی‌ترین روش‌های کاهش ابعاد داده‌ها در مدل‌های چند متغیره است. مؤلفه‌های اصلی با توجه به خصوصیاتی که دارند برای مقابله با مشکل هم خطی چندگانه و کاهش ابعاد داده‌ها مورد استفاده قرار می‌گیرند. در این روش با استفاده از ماتریس مقادیر ویژه، مؤلفه‌های اصلی به صورت ترکیبی خطی از متغیرهای اولیه و مستقل از یکدیگر ساخته شده و در تحلیل داده‌ها به جای متغیرهای اولیه مورد استفاده قرار می‌گیرند.

در این تحقیق برای اعمال روش PCA از کتابخانهٔ Scikit Learn که از کتابخانه‌های متن‌باز، مفید، پرکاربرد و قدرتمند در زبان برنامه‌نویسی پایتون است و برای اهداف یادگیری ماشین به کار می‌رود، استفاده شده است. هدف، پیدا کردن مؤلفه‌هایی است که بیشترین واریانس را نشان می‌دهند. این رویکرد به این دلیل است که در روش تحلیل مؤلفه اصلی، هدف، پیدا کردن مؤلفه‌هایی است که بیشترین اطلاعات را دربر دارند. کد نوشته شده در این پژوهش ۴۵۹ مولفهٔ اصلی را به عنوان خروجی ارائه می‌کند.

نتایج و بحث

مجموعه داده‌ها

متغیرهای مورد بررسی در این تحقیق، کلماتی هستند که بار معنایی مثبت، منفی و یا خنثی‌ای دارند. این متغیرها از داده‌های مورد بررسی در این پژوهش که داده‌های متنی حاوی نظرات بیمه‌گذاران بیمهٔ زندگی یک شرکت بیمه ای می‌باشد به دست آمده اند. این نظرات با جمع‌آوری از سامانهٔ شرکت، فرم‌های نظرسنجی، فرم‌های اندازه‌گیری سطح رضایت مشتریان، نظرات درج شده در قسمت امور مشتریان شرکت و مصاحبه‌های حضوری که با برخی از بیمه‌گذاران صورت گرفته گردآوری شده است. تعداد کل این نظرات ۱۰۰۰۰ (ده هزار) کامنت و به صورت جملات محاوره‌ای و غیر رسمی بوده است که در بازهٔ زمانی اول فروردین ۱۴۰۰ الی انتهای شهریور ۱۴۰۰ بیان شده است.

آماده‌سازی داده‌ها

با گسترش سیستم‌های پایگاهی و حجم بالای داده‌های ذخیره شده دراین سیستم‌ها، به ابزاری نیاز است که بتوان این داده‌ها را پردازش کرده و اطلاعات حاصل از آن را دراختیار کاربران قرار داد. فرایند داده کاوی این وظیفه را برعهده دارد، اما دستیابی به نتایج واقعی و مؤثر بدون برخورداری از ورودیهای صحیح و قابل اعتماد ممکن نیست، لذا می بایست پیش از هرتحلیلی ازصحت و تناسب داده‌ها و اطلاعات موجود اطمینان داشته باشیم (Motevali Haghi *et al.*, 2012). از این رو، در این پژوهش نیز از تعداد نظرات ذکر شده، با سپری‌شدن مرحلهٔ پیش پردازش، به تعداد ۲۰۰۰ کامنت رسیدیم. در نهایت با در نظر گرفتن درصد ۸۰ و ۲۰ برای آموزش و آزمایش، تعداد ۴۰۰ نظر برای آزمایش کنار گذاشته شد و نسبت به برچسب زنی ۱۶۰۰ کامنت باقی مانده که با رعایت امر توازن، شامل ۸۰۰ نظر مثبت و ۸۰۰ نظر منفی بود، با قصد طبقه‌بندی کردن نظرات به دو کلاس

صفر و یک (نظرات منفی و مثبت) اقدام شد. برای برچسب‌گذاری نظرات از یک طیف لیکرت ۵ گزینه‌ای شامل: ۱ (رضایت بسیار کم)، ۲ (رضایت کم)، ۳ (رضایت متوسط)، ۴ (رضایت بالا) و ۵ (رضایت خیلی بالا) استفاده شده است. به این صورت که نظرات با برچسب ۱ و ۲ در دسته بندی نظرات منفی و نظرات با برچسب بالاتر در دسته بندی نظرات مثبت قرار خواهند گرفت.

این برچسب‌گذاری به صورت دستی و توسط خبرگان این صنعت صورت گرفته است. تعداد این خبرگان ۱۰ نفر و شامل اساتید دانشگاه در رشتهٔ مدیریت بیمه (۵ نفر)، مدیران میانی شرکت بیمهٔ مورد بررسی (۳ نفر) و (۲) نفر از مدیران امور مشتریان شرکت بوده و تعداد نظرات به صورت یکسان برای برچسب‌گذاری بین این افراد تقسیم گردیده است (برای هر فرد خبره ۱۰۰۰ نظر).

نمونه‌ای از داده‌های برچسب گذاری شده، در جدول ۱ مشاهده می‌شود.

اطلاعات دموگرافیک در نظر گرفته برای این پژوهش شامل ۲۰ متغیر به شرح جدول ۲ بوده است:

نسبت تعداد نظرات مثبت و منفی در شکل ۲ نشان داده شده است:

روش‌های ارزیابی مدل

در این پژوهش، برای ارزیابی و مقایسهٔ مدل‌ها از معیارهای دقت، بازخوانی، صحت و معیار اف استفاده کرده‌ایم:

دقت: اولین معیار یا سنجه‌ای که به ذهن می‌رسد، معیار دقت یا میزان تشخیص درست مدل است؛ یعنی نسبت تشخیص‌های درست (مثبت واقعی+ منفی واقعی) به کل داده‌ها. این معیار برای ارزیابی مدل‌ها در زمان استفاده از داده‌های نامتوازن (یعنی تفاوت زیادی در تعداد نمونه‌های دسته‌ها وجود دارد) کافی نیست؛ زیرا این عدم توازن باعث می‌شود مدل‌های متمایل به دستهٔ پرتعداد، شناسایی نشوند و به اشتباه یک مدل بد، مدل خوبی معرفی شود.

بازخوانی: ارزیابی این که کل نمونه‌های واقعاً مثبت شامل نمونه‌هایی است که درست، مثبت شناسایی شده‌اند (مثبت واقعی) و نمونه‌هایی که مثبت بوده‌اند اما نادرست، منفی شناسایی شده‌اند (منفی اشتباه). در این معیار، بر تعداد نمونه‌های مثبت شناسایی شده به کل نمونه‌های مثبت تمرکز می‌شود.

صحت: در کنار معیار بازخوانی معیار دیگری به نام صحت، که برابر تعداد نمونه‌های تشخیصی مثبت واقعی به کل نمونه‌های مثبت اعلام شده است تعریف می‌شود تا میزان مثبت‌های اشتباه هم در نظر گرفته شود.

معیار اف: «میانگین هارمونیک بازخوانی و صحت می‌باشد» (Han *et al.*, 2012).

پیاده‌سازی مدل

مراحل انجام شده در این پژوهش، به این شرح صورت گرفت که پس از جمع‌آوری داده‌ها، مرحلهٔ پیش پردازش و در اولین قدم این مرحله استخراج توکن‌ها شکل گرفت تا بتوانیم بر روی برداری از کلمات کار کنیم.

حذف نمادها و کاراکترهای ویژه، حذف نیم فاصله‌ها، حذف تک کاراکترها، جایگزینی چند فاصلهٔ خالی با یک فاصلهٔ خالی، حذف کلمات ایستا، جایگزینی کلمات محاوره‌ای با کلمات رسمی، استخراج ریشهٔ لغات، برچسب‌گذاری نقش کلمات، اعمال روش TFIDF، اعمال روش PCA و تقسیم داده‌ها به آموزش و آزمایش به نسبت ۸۰ به ۲۰ از جمله کارهای دیگر انجام شده در این پژوهش بود.

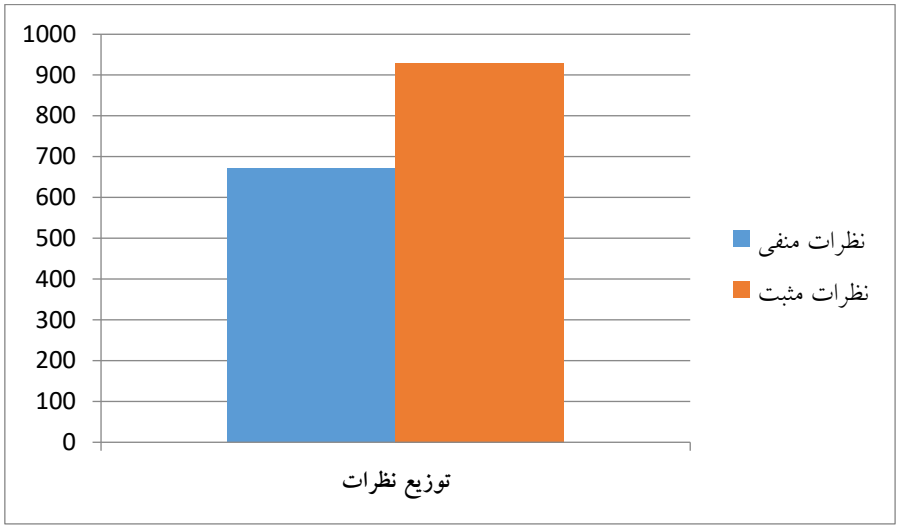
از میان الگوریتم‌های ماشین، چهار الگوریتم ماشین بردار پشتیبان، جنگل تصادفی، حافظهٔ طولانی کوتاه ـ مدت و بیزین ساده در این پژوهش مورد استفاده قرار گرفته است. دلیل استفاده از این الگوریتم‌ها، استفاده از آنها در پژوهش‌های مختلفی بوده که غالباً

جدول ۱: نمونه‌ای از داده‌های برچسب‌گذاری شده

۱	حقوق بازنشستگی بیمهٔ زندگی، با توجه به نرخ تورم خیلی کم است.
۵	من از طریق بیمهٔ زندگی وام گرفتم. راضی هستم. نه ضامن خواست نه هیچی.
۴	بیمهٔ زندگی برای امیدواری آدم نسبت به آینده خوبه. من خودمو تحت پوشش بیمه زندگی درآوردم. از اون موقع هم آرامشم بیشتر شده و هم امید به زندگیم.
۵	بیمهٔ زندگی نسبت به بقیهٔ شرکت‌های بیمه‌گر، اعتبار بالایی داره.
۱	کارشناسای بیمهٔ زندگی شرکت خدماتشون رو خیلی نامناسب ارائه میدن انگار از آدم طلبکارند.
۵	امکان بازخرید، امکان خیلی خوب بیمه زندگی هست.
۱	تو بیمهٔ زندگی مقدار حقوق بازنشستگی نسبت به پولی که میگیرن خیلی کمه.
۳	بیمه زندگی خیلی خوبه. حقوق بازنشستگیش هر چقدرم کم باز ضروریه. به آدم قوت قلب و اطمینان خاطر میده.
۵	من که از خدمات بیمهٔ زندگی راضی هستم. راحت و بی‌دردسر و بدون اتلاف وقت غرامت بیماریموگرفتم.
۲	من بیشتر کارهام رو اینترنتی انجام میدم. یه بار که برای سؤالی دربارهٔ بیمهٔ زندگی مراجعه کرده بودم تو مراجعهٔ حضوری برخورد تقریباً بدی از طرف کارشناساش دیدم.
۳	حقوق بازنشستگی بیمهٔ زندگی نسبت به تورم تقریباً بد نیست.
۵	بیمهٔ زندگی خیلی خوبه. به نظر من هر شهروندی باید یه بیمه‌نامهٔ زندگی بگیره تا دیگه دغدغهٔ آینده رو نداشته باشه.
۲	فقط طرح وامشون تو بیمهٔ زندگی طرح خوبیه که ضامن نمیخواد وگرنه بقیهٔ چیزاش شبیه کلاهبرداریه.
۱	کسی که اقتصاد مهندسی خونده باشه میفهمه که بیمهٔ زندگی برای سرمایه‌گذاری خیلی بده. اصلاً به این بیمه به چشم سرمایه‌گذاری نباید نگاه کرد.

جدول ۲: لیست متغیرها

ردیف	متغیرها
۱	جنسیت
۲	وضعیت تأهل
۳	بیمه‌گذار
۴	سن بیمه شده
۵	تعداد سالی که مشتری شرکت بیمه است ؟
۶	پوشش اضافی بیمهٔ زندگی
۷	درآمد ماهانهٔ بیمه‌گذار
۸	روش پرداخت حق بیمه
۹	برداشت از محل اندوختهٔ بیمه‌نامهٔ زندگی صورت گرفته است؟
۱۰	بیمه‌گذار از بیماری خاصی رنج می‌برد؟
۱۱	بیمه‌گذار از پوشش بیمهٔ زندگی و مستمری طرح خانوادهٔ بیمهٔ زندگی استفاده کرده است؟
۱۲	سابقهٔ دریافت خسارت یا غرامت بر اساس پوشش اصلی یا پوشش‌های اضافی بیمهٔ زندگی (به‌دلیل امراض یا فوت)
۱۳	آیا بیمه‌گذار بیمه‌نامهٔ درمان تکمیلی انفرادی متصل به زندگی را نیز تهیه نموده است؟
۱۴	بیمه‌گذار از چند طرح بیمهٔ زندگی استفاده می‌کند؟
۱۵	سابقهٔ دریافت خسارت یا غرامت فقط به‌دلیل حادثه
۱۶	چند طرح از طرح‌های پوشش امراض خاص توسط بیمه‌گذار تهیه شده است؟
۱۷	به غیر از این شرکت، بیمه‌گذار از پوشش یا پوشش‌های بیمهٔ زندگی بیمه‌گران دیگر نیز استفاده می‌کند؟
۱۸	بیمه‌شده با توجه به های ریسک بودن از نظر بیماری، در لیست سیاه شرکت قرار دارد؟
۱۹	آیا بیمه‌گذار به غیر از بیمهٔ زندگی از رشته‌های بیمه‌ای دیگر شرکت نیز استفاده می‌کند؟
۲۰	آیا سابقه‌ای از شکایت بیمه‌گذار در واحد CRM شرکت موجود است؟



شکل ۲: توزیع نظرات

این الگوریتم‌ها نتایج قابل‌قبولی را حاصل نموده‌اند. در به کارگیری این الگوریتم‌ها به طور معمول از پارامترهای پیش فرض آنها استفاده شده است. به عنوان نمونه، در روش ماشین بردار پشتیبان از SVM غیرخطی و با kernel=rbf استفاده شده است و یا در روش LSTM از یک لایه Emmbeding به اندازهٔ ۱۰۰، یک لایه Dropout، یک لایه LSTM با ۱۰ نورون و یک لایهٔ انتهایی Dense با ۲ نورون و تابع فعال سازی Softmax استفاده شده است.

خلاصه‌ای از نتایج این آزمایش‌ها در جدول ۳ ارائه شده است. در جدول ۳، نتایج معیارهای صحت، بازخوانی و معیار اف برای هر دو طبقهٔ صفر و یک (نظرات منفی و مثبت) بیمه‌گذاران بیمه زندگی ارائه شده است:

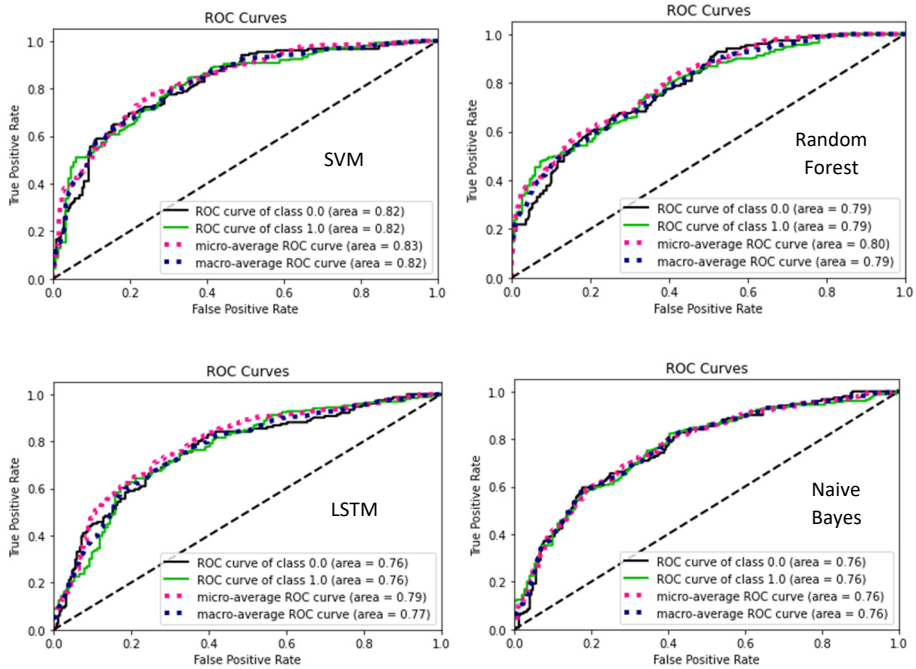
همین طور که مشاهده می‌شود برای هر دو طبقهٔ مثبت و منفی،

جدول ۳: نتایج

الگوریتم	کلاس صفر			کلاس ۱			دقت
	صحت	بازخوانی	معیار اف	صحت	بازخوانی	معیار اف	
بیزین ساده	۰/۷۴	۰/۳۲	۰/۴۴	۰/۶۹	۰/۹۳	۰/۷۹	۰/۷
جنگل تصادفی	۰/۸	۰/۲۴	۰/۳۷	۰/۶۸	۰/۹۶	۰/۷۹	۰/۶۹
ماشین بردار پشتیبان	۰/۷۵	۰/۴۱	۰/۵۳	۰/۷۲	۰/۹۲	۰/۸۱	۰/۷۳
حافظه طولانی کوتاه- مدت	۰/۴۱	۰/۴۱	۰/۴۱	۰/۷۲	۰/۷۲	۰/۷۲	۰/۶۲

جدول ۴:. ماتریس سردرگمی الگوریتم ماشین بردار پشتیبان

	۰	۱	جمع
۰	۶۲	۸۹	۱۵۱
۱	۲۱	۲۲۸	۲۴۹
جمع	۸۳	۳۱۷	۴۰۰



شکل ۳ : نمودارهای مشخصهٔ عملکرد

معیارها محاسبه شده و در نهایت برای هر الگوریتم نیز معیار دقت تعیین شده است. با توجه به نتایج حاصله می‌توان گفت که در کلاس صفر، الگوریتم جنگل تصادفی در معیار صحت و الگوریتم ماشین بردار پشتیبان در معیارهای بازخوانی و معیار اف بهترین نتایج را داشته‌اند. در کلاس ۱ نیز، الگوریتم ماشین بردار پشتیبان در معیارهای صحت و معیار اف و الگوریتم جنگل تصادفی در معیار بازخوانی بهترین نتایج را کسب کرده‌اند. همچنین الگوریتم ماشین بردار پشتیبان از منظر دقت در هر دو کلاس صفر و یک بیشترین میزان را داشته است. یکی از رایج ترین مسائل در مدل‌های یادگیری با ناظر طبقه‌بندی داده‌ها است. بدیهی است که این گونه از مدل‌ها نیز نیاز به معیارهایی برای ارزیابی دارند. مهمترین معیار ارزیابی این مدل‌ها ماتریس سردرگمی است(Fadaei eslam, 2021). جهت ارائهٔ جزئیات بیشتر، ماتریس سردرگمی الگوریتم ماشین بردار پشتیبان در جدول ۴ ارائه شده است.

یکی دیگر از روش‌های بررسی و ارزیابی عملکرد الگوریتم‌های دسته‌بندی، نمودار مشخصهٔ عملکرد است. این نمودار توسط ترسیم

نسبت نرخ مثبت صحیح برحسب نرخ مثبت کاذب ایجاد می‌شود (Kiadaliri and Azizi, 2020). در شکل ۳، این نمودار برای الگوریتم‌های ماشین بردار پشتیبان، جنگل تصادفی، بیزین ساده و حافظهٔ طولانی کوتاه مدت ارائه شده است.

به طور کلی، با شرط مشخص بودن توزیع احتمالی برای هر دو بخش نرخ مثبت صحیح و نرخ مثبت کاذب، منحنی مشخصهٔ عملکرد در صورتی حاصل خواهد شد که تابع توزیع تجمعی یا سطح زیر منحنی توزیع احتمال تشخیص درست را در محور عمودی و تابع توزیع تجمعی تشخیص نادرست را در محور افقی در نظر بگیریم. با توجه به در نظر گرفتن سطح زیر منحنی در نمودارها، الگوریتم ماشین بردار پشتیبان در هر دو کلاس صفر و یک بهترین عملکرد را داشته و الگوریتم جنگل تصادفی از این منظر در ردهٔ بعدی قرار می‌گیرد.

جمع‌بندی و پیشنهادها

الگوریتم‌های به کار گرفته شده در این پژوهش، به منظور تحلیل نظرات کاربران بیمه‌گذار استفاده کننده از خدمات بیمهٔ زندگی یک شرکت بیمه ای در استان تهران بوده است. با توجه به حجم و لحن عبارات به کار برده شده توسط بیمه‌گذاران، مزیت استفاده از این شیوه قابل درک است. تکنیک‌های مختلفی برای یادگیری ماشین استفاده می‌شوند. تکنیک‌های با ناظر، بدون ناظر، شبه ناظر و تقویتی از پرکاربردترین روش‌های یادگیری ماشین هستند که از روش‌های طبقه بندی، رگرسیون، خوشه بندی، قوانین انجمنی، کاهش ابعاد و سایر روش‌ها برای یادگیری استفاده می‌کنند (Pahlevani Ghomi *et al.*, 2020). فرایند یادگیری در ساختار این پژوهش از زیر مجموعهٔ طبقه‌بندی همراه با ناظر است. به‌طور کلی می‌توان نتیجه گرفت که در این پژوهش به کارگیری الگوریتم‌های مختلف داده کاوی، نتایج مختلفی را از منظر معیارهای مختلف حاصل کرده است. همان‌طور که در بخش پیشین اشاره شد، در کلاس صفر الگوریتم جنگل تصادفی در معیار صحت و الگوریتم ماشین بردار پشتیبان در معیارهای بازخوانی و اف بهترین نتایج را داشته‌اند. در کلاس ۱ نیز، الگوریتم ماشین بردار پشتیبان در معیارهای صحت و اف و الگوریتم جنگل تصادفی در معیار بازخوانی بهترین نتایج را کسب کرده‌اند. همچنین الگوریتم ماشین بردار پشتیبان از منظر دقت در هر دو کلاس صفر و یک بیشترین میزان را داشته است.

با استفاده از نتایج این پژوهش می‌توان گفت که اکثریت بیمه‌گذاران نظر مثبتی در ارتباط با خدمات دریافتی داشته‌اند و مایلند در آینده نیز این خدمت بیمه‌ای را در سبد خرید خود داشته باشند. به طور دقیق تر با توجه به متغیرهای تحقیق در این پژوهش، نتیجه گیری می‌شود که مشتریان مرد متأهل با درآمد بالای ۱۵ میلیون تومان در ماه که هیچ سابقهٔ شکایتی از آنها در واحد امور مشتریان شرکت وجود ندارد، بیشترین سهم را در میزان نظرات مثبت داشته اند. لذا مسئولین بیمه می‌توانند مشتریان احتمالی آتی خود را از میان این افراد پیدا و با تعیین یک استراتژی بازاریابی مناسب، برای فروش خدمات خود بر روی آنها سرمایه‌گذاری کنند. به عنوان پیشنهاداتی برای تحقیقات آتی، می‌توان به افزودن متغیرهای دیگر به لیست متغیرها، استفاده از Kernel های دیگر در روش ماشین بردار پشتیبان و یا Emmbeding متفاوت در روش LSTM و یا آزمایش تکنیک کشف احساسات از روی چهره اشاره کرد. در آخر باید اضافه کرد که شرکت‌های بیمه علاوه بر در پیش گرفتن روش‌های نوین بازاریابی برای جذب مشتریان جدید، باید همزمان به پیش‌بینی سبد خرید مشتریان خود نیز بپردازند. طراحی سیستم‌های تصمیم یار هوشمند برای این نوع پیش‌بینی می‌تواند موضوعی قابل توجه برای تحقیقات آتی تلقی شود.

مشارکت نویسندگان

علیرضا امین پور: نگارش اولیه، جمع آوری مطالب و پیاده سازی، محمد ربیعی: نظارت بر حسن اجرا، اصلاح و جمع‌بندی.

تشکر و قدردانی

این مقاله مستخرج از رسالهٔ دکتری علیرضا امین پور با عنوان «طراحی سیستم های تصمیم یار هوشمند برای پیش بینی سبد خرید مشتریان خدمات بیمه با استفاده از کلان داده» دانشگاه ایوان کی و با راهنمایی دکتر محمد ربیعی می‌باشد؛ بدین وسیله از راهنمایی‌ها و مشاوره‌های ایشان تشکر ویژه می نمایم.

تعارض منافع

نویسنده(گان) اعلام می‌دارند که در مورد انتشار این مقاله تضاد منافع وجود ندارد. علاوه بر این، موضوعات اخلاقی شامل سرقت ادبی، رضایت آگاهانه، سوءرفتار، جعل داده‌ها، انتشار و ارسال مجدد و مکرر توسط نویسندگان رعایت شده است.

دسترسی آزاد

یادداشت ناشر

منابع

Amarasinghe, H.; Warnakulasuriya, S.; Johnson, N. W., (2021). Evaluation of a social marketing campaign for the early detection of oral potentially malignant disorders and oral cancer: Sri Lankan experience. J. Oral Biol. Cran. Res., 11(2): 204-208 **(5 Pages).**

Asghari, R., (2012). Application of N-gram model in statistical language modeling. International conference on nonlinear modeling and optimization **(7 Pages).** [In Persian]

Babania, M.; Pourdarvish, A.; Mirashrafi, S., (2020). The role of principal component analysis (PCA) in big data modeling. The third international conference on soft computing **(9 Pages).** [In Persian]

Berson, A.; Smith, S. J.; Thearling, K., (1999). Building data mining applications for CRM. McGraw-Hill**.**

Boroun, G.; Raad, F.; Parvin, H., (2015). Evaluation analysis and combination of machine learning algorithms and data mining techniques for classification. The third international conference on applied research in computer engineering and information technology **(13 Pages).** [In Persian]

Cheo Yeo, A.; Smith, A.K.; Willis, R.J.; Brooks, M., (2001). Modeling the effect of premium changes on motor insurance customer retention rates using neural networks. International Conference on Computational Science: 390-399 **(10 Pages).**

Fadaei eslam, M.J., (2021). Co-partitioning using the two-stage structured matrix analysis method. Mach. Vision Image Process., 7(2): **(11 Pages).** [In Persian]

Farokhseir, G.; Esmaeelpour, M., (2015). Improvement of user behavior pattern recognition algorithms in web mining using link analysis. Second international conference on electrical engineering and computer science **(8 Pages).** [In Persian]

Gharakhani, M.; Abolghasemi, M., (2011). Applications of data mining in the insurance industry. J. News. World Insur., 6(158): 5-21 **(17 Pages).** [In Persian]

Hajiheydari, N.; Khaleh, S.; Farahi, A., (2011). Classification of the risk level of car body insurance policyholders using data mining algorithms (Case study: An insurance company). Iran. J. Insur. Res., 26(4): 107-129 **(23 Pages).** [In Persian]

Han, J.; Kamber, M.; Pei, J., (2012). Data mining concepts and techniques (Third edition). USA: Elsevier**.**

Homburg, C.; Jozić, D.; Kuehnl, C., (2017). Customer experience management: Toward implementing an evolving marketing concept. J. Acad. Marketing. Sci., 45(3): 377-401 **(25 Pages).**

Hsia, J.; Kemper, E.; Kiefe, C.; Zapka, J.; Sofaer, S.; Pettinger, M.; Bowen, D.; Limacher, M.; Lillington, L.; Mason, E., (2000). The importance of health insurance as a determinant of cancer screening: Evidence from the Women's Health Initiative. Preventive. Med., 31(3): 261-270 **(10 Pages).**

Izadparast, M.; Farahi, A.; Fathnezhad, F.; Teimourpour, B., (2022). Using data mining methods to predict the damage level of car body insurance customers. J. Inf. Process. Manage., 27(3): 699-722 **(24 Pages).** [In Persian]

Jun mei, D.; Gui Quan, L.; Hui, L., (2015). The Application of Improved Random Forest in the Telecom Customer Churn Prediction. J. Telecom. Res. App, 28(11): 1041-1049 **(9 Pages).**

Kiadaliri, F.; Azizi, A., (2020). An overview of performance evaluation indicators of control charts. The first international conference on new challenges and solutions in industrial engineering and management and accounting **(10 Pages).** [In Persian]

Manteghipour, M.; Alaei, M., (2022). Discount effects on the composition of the risk portfolio of the third-party vehicle insurance. Iran. J. Insur. Res., 10(2): **(28 Pages).** [In Persian]

Momeni, F.; Ghodousi, M., (2016). Decision support systems and business intelligence and design and implementation of a decision support system (case study of insurance industry). The 4th International conference on information, communication and computer technology **(11 Pages).** [In Persian]

Motarjem, K.; Niakan, L., (2020). Measuring and evaluating the satisfaction of life insurance customers. Iranian J., 10(1): 87-119 **(33 Pages).** [In Persian]

Motevali Haghi, M.; Ahmadian Tabasi, H.; Sajadi, N., (2012). Data preparation for data mining techniques and its application. The first national conference of information technology and computer networks of PNU **(16 Pages)**. [In Persian]

Oliaei, A.; Hazeri, A.; Jamali, M.; Khazaee, A., (2018). Finding potential customers to buy insurance with data mining techniques. The second national conference of computer, information technology and artificial intelligence applications **(7 Pages).** [In Persian]

Oshini, G. T. L.; Caldera, H. A., (2013). Mining life insurance data for customer attrition analysis. J. Indust. Intellig. Inf., 1(1): 52-58 **(7 Pages).**

Pahlevani Ghomi, M.; Osati Araghi, N.; Nazari, S., (2020). Machine learning techniques and algorithms. The second national conference of new skills in electrical engineering, computer and communication technology **(7 Pages).** [In Persian]

Palakshappa, A.; Patil, M.M., (2022). RFM model for customer purchase behavior using K-Means algorithm. J. King Saud Univ. Comput. Inf. Sci., 34(5): 1785-1792 **(8 Pages).**

Rezaei navaei, S.; Koosha, H., (2017). Applying and evaluating data mining techniques to predict customer turnover in the insurance industry. Int. J. Ind. Eng. Oper. Manage., 27(4): 635-653 **(19 Pages).** [In Persian]

Senousy, Y.; Nashaat El-Khamisy, M.; Alaa El-din, R., (2018). Recent trends in big data analytics towards more enhanced insurance business models. Int. J. Comp. Sci Inf. Secur., 16(12): 39-45 **(7 Pages).**

Sagiroglu, S.; Sinanc, D., (2013). Big data: A review. Collaboration technologies and systems (CTS). 2013 International conference on collaboration technologies and systems.

Turney, P.D., (2002). Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. arXiv preprint cs/0212032.

Vali Mohammadi, S.; Shekarchi, A., (2010). The application of data mining in improving customer relationship management. The first national industrial and systems engineer-

ing conference **(11 Pages).** [In Persian]
Yoon, S. J.; Lee, H.J., (2017). Does customer experience management pay off? Evidence from local versus global hotel brands in South Korea. J. Hospitaly marketing manage., 26(6): 585-605 **(21 Pages).**
Zaresaadabadi, E.; Zaresaadabadi, M., (2014). Data mining is a new way to face challenges. The third national accounting and management conference **(16 Pages).** [In Persian]

AUTHOR(S) BIOSKETCHES **معرفی نویسندگان**

علیرضا امین پور، دانشجوی دکترای رشتهٔ مهندسی صنایع، گروه مهندسی صنایع، دانشکده فنی و مهندسی، دانشگاه ایوان کی، ایوان کی، ایران

- Email: *Aminpour_A@eyc.ac.ir*
- ORCID: 0000-0003-0243-4067
- Homepage: https://www.eyc.ac.ir/ShowMainPage2.aspx?mid=940

محمد ربیعی، استادیار و عضو هیئت علمی، گروه مهندسی برق و کامپیوتر، دانشکده فنی و مهندسی، دانشگاه ایوان کی، ایوان کی، ایران

- Email: *mohammad.rabiei@uniud.it*
- ORCID: 0000-0002-5430-3846
- Homepage: https://www.eyc.ac.ir/ShowProfessor.aspx?mid=29519

HOW TO CITE THIS ARTICLE

Aminpour, A.; Rabiei, M., (2023). Classification of life insurance customer point of views based on text mining algorithms. Iran. J. Insur. Res., 12(1): 15-26.

DOI: 10.22056/ijir.2023.01.02

URL: https://ijir.irc.ac.ir/article_159477.html?lang=en